Recommending System for Opening Businesses in Bogota, Colombia

1. Introduction

1.1. Background

Due to the Covid Pandemic that rapidly spread all over the world in the year 2020, various restrictions were put in place in order to slow down the spreading of the virus. Bogota, Colombia was one of the cities that entered into a lockdown in March 2020, and due to this, various businesses had to shut down because of the lack of clients and people began losing their jobs. Now that some restrictions are being lifted in 2021, people are considering reopening their businesses or even creating new ones.

1.2. Problem

In order to open a business, people would be interested in a solution that helps them determine the best location by taking into account multiple factors such as: Trending venues in the localities, information on similar localities access to outdoor spaces (a plus in order to maintain the covid protocols established by the government) and covid cases in the locality. These are factors that must be taken into account at the moment because a locality with a peak in covid cases and a lack of covid protocol could be sent into lockdown again and impact the new business.

1.3. Possible Stakeholders

This project or recommendation system would be of interest to those persons who want to open a business and want to determine which are some of the best location options taking into account the factors mentioned previously. For this version of the project the focus will be on stakeholders who would like to open businesses in the categories of bookstores, coffee shops or book cafés.

2. Data

2.1. Data Sources

2.1.1. Bogota Localities and Neighbourhoods

We will be needing information about Bogota´s localities and the neighbourhoods within these localities. This information can be found on this website. In here we can find the neighbourhood id, its name and the locality it belongs to.

We can visualize an example of how the website supplies the data on the next figure:

Número	Nombre	Localidad
1	Paseo de los Libertadores	
9	Verbenal	
10	La Uribe	
11	San Cristóbal Norte	
12	Toberín	01 Usaquén
13	Los Cedros	
14	Usaquén	
15	Country Club	
16	Santa Bárbara	

Figure 1. Preview of neighbourhoods and localities data

2.1.2. FourSquare API / Geopy Library

The Geopy Library is used to get the coordinates of the neighborhoods in Bogota so that then these can be used as an input for the FourSquare API. The coordinates were then added to the neighborhoods dataframe.

By having the location information of Bogota's localities, the FourSquare API can be used to get information on the venues located within these localities. Depending on the kind of request sent to the API more information on the venues can be returned. For the recommender system in the project the information which could be considered valuable is: Name, Location, Category, Rating.

Since the focus of this version of the project is businesses in the category of coffee shops, bookstores and book cafés, the FourSquare API will be called using the additional filter of CategoryId. The categories used were:

- Café - `4bf58dd8d48988d16d941735`
- Bookstores - `4bf58dd8d48988d114951735`
- Coffee Shops - `4bf58dd8d48988d1e0931735`
- Parks - `4bf58dd8d48988d163941735`

	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Id
	Juan Valdez Café	4.762551	-74.045620	Café	4bf58dd8d48988d16d941735
	Crepes & Waffles	4.762301	-74.045801	Café	4bf58dd8d48988d16d941735
	Juan Valdez Café	4.762868	-74.045356	Café	4bf58dd8d48988d16d941735
	Librería Nacional, Santafé Mi Mundo - Bogotá	4.761717	-74.045916	Bookstore	4bf58dd8d48988d114951735
	Artesanos	4.762598	-74.045745	Café	4bf58dd8d48988d16d941735

Figure 2. Preview of a FourSquare API response (stored in a dataframe)

2.1.3. Bogota Covid Cases by Localities

In order to determine which localities have a high probability of going into lockdown again or to implement stronger restrictions, it is necessary to have recent data on the amount of covid cases in the localities in Bogota. This information can be found in different Municipal Government websites, from which we'll be using this one. On this website a database of covid cases in Bogota can be downloaded and used for the recommender system. This database includes the following information: Case number, Date of the initial symptoms, Date of the diagnosis, City, Locality, Age, Gender, Contagion source, State of the patient. This database has over 1.2M records.

_id	CASO	FECHA_...	FECHA_...	CIUDAD	LOCALI...	EDAD	UNI_MED	SEXO	FUENTE...	UBICAC...	ESTADO
292359	292359	2020-10-...	2020-10-...	Bogotá	Engativá	43	1	F	En estudio	Casa	Recuper...
292360	292360	2020-10-...	2020-10-...	Bogotá	Suba	41	1	F	En estudio	Casa	Recuper...
292361	292361	2020-10-...	2020-10-...	Fuera de...	Fuera de...	41	1	F	En estudio	Casa	Recuper...
292362	292362	2020-09-...	2020-10-...	Bogotá	Suba	40	1	F	En estudio	Casa	Recuper...
292363	292363	2020-10-...	2020-10-...	Bogotá	Suba	40	1	F	En estudio	Casa	Recuper...
292364	292364	2020-10-11	2020-10-...	Bogotá	Engativá	40	1	F	Relacion...	Casa	Recuper...
292365	292365	2020-09-...	2020-10-...	Bogotá	Suba	45	1	F	Relacion...	Casa	Recuper...
292366	292366	2020-10-11	2020-10-...	Bogotá	Usaquén	45	1	F	Relacion...	Casa	Recuper...

Figure 3. Bogota Covid cases database preview

2.2. Data Cleaning / Data Enrichment

2.2.1. Neighborhoods Dataframe

Since the table that contains the data needed doesn't have a structure that can be easily read into a pandas dataframe, the data had to be downloaded, the localities had to be ungrouped and since we only need the name of the locality, we had to split the values and keep only the name and not the number that accompanied this column.

A function was created in order to get the coordinates of each neighborhood (by using the geopy libraries) and add them to this neighborhood dataframe.

Seven neighborhoods weren't found using the geopy libraries so the coordinates had to be added manually.

	Number	Neighborhood	Locality	Latitude	Longitude
0	1	Paseo de los Libertadores	usaquén	0.000000	0.000000
1	9	Verbenal	usaquén	4.765150	-74.038394
2	10	La Uribe	usaquén	4.752400	-74.045013
3	11	San Cristóbal Norte	usaquén	4.734501	-74.017543
4	12	Toberín	usaquén	4.747274	-74.043719

Figure 4. Preview of the neighborhoods dataframe

2.2.2. Bogota Covid Cases Dataframe

For this dataset, the data cleansing process consisted of various steps:

- We can see that some of the records are from outside of Bogota (in the 'CIUDAD' column) or do not have information about the city. Since Bogota is our target city for this project, we are going to delete these records
- The dataset contains various columns which have information that are not relevant to the project, that's why the following columns were dropped: CASO, FECHA_DE_INICIO_DE_SINTOMAS, CIUDAD, EDAD, UNI_MED, SEXO, FUENTE_O_TIPO_DE_CONTAGIO, UBICACION, ESTADO. In English: Case number, symptoms initial date, city, age, uni_med, gender, type of contagion, location (house/hospital), status
- The remaining columns consist of FECHA_DIAGNOSTICO, LOCALIDAD_ASIS. In English: Diagnosis date, locality. The diagnosis date had to be cast as a date format since originally it was cast as an object format.
- Create a new column where only the month and year of the diagnosis date will be stored

- Group and count the values in the dataframe so that it shows information on the amount of covid cases diagnosed in a time period (month/year) in each Locality

	LOCALIDAD_ASIS	month_year	Cases Diagnosed
0	antonio nariño	2020-03-01	4
1	antonio nariño	2020-04-01	26
2	antonio nariño	2020-05-01	163
3	antonio nariño	2020-06-01	418
4	antonio nariño	2020-07-01	1221

Figure 5. Covid cases dataframe

2.2.3. Matching the datasets

Since later on in the project we will need to match the localities in the neighborhoods and covid dataset, it is necessary to have the same standard for the localities names and check if we have relevant data for all of the localities listed. The following transformations and updates were made on the data:

- Changed all localities names to lowercase format
- Changed 'rafael uribe' to 'rafael uribe uribe' in the neighbourhoods dataframe
- Changed 'mártires' to 'los mártires' in neighbourhoods dataframe
- Deleted records with the sumapaz locality since, unfortunately, we don't have enough relevant information from this locality

Some of the information that was downloaded was in spanish. In order to avoid merging conflicts and because the main language of the course/project is english; the column names of these data frames were translated.

3. Data Exploration

3.1. Bogota's Neighborhoods Map

Once the neighborhoods dataframe contained information on the name of the neighborhoods, their corresponding locality and their coordinates: by using the Folium Python Library a map of the city was built in order to have a better understanding of the city and the neighborhoods locations.

Figure 6. Map of Bogota´s Neighborhoods

3.2. Bogota's Venues

Based on the target venues dataframe that was built with the information returned from the FourSquare API, by using the Folium Python Library a map will be built which includes markers for the different venues and its category.
Red - Coffee Shops / Cafés
Blue - Book shops
Green - Parks

According to this map there is a large concentration of our target venues in the eastern side of the city. This observation might be useful for later on when we start our clustering process for the neighborhoods.

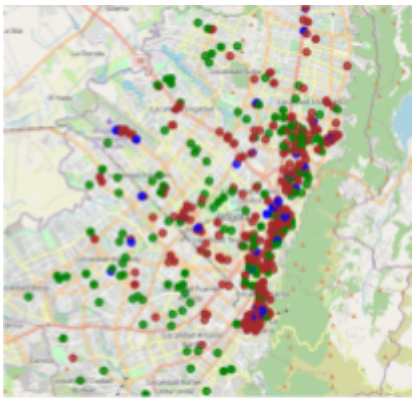


Figure 7. Map of Bogota's target venues

3.3. Covid Data

The covid data that was downloaded includes records from March/2020 to our current date (At the moment of this report, July 20.2021). In order to visualize the behaviour in the amount of covid cases diagnosed per locality, we will be using the Seaborn Python Library for bar graph generation.

Before building the graphs we use the pandas describe function to get a basic understanding of the data.

	Cases Diagnosed
count	333.000000
mean	3626.603604
std	4982.653493
min	1.000000
25%	542.000000
50%	2006.000000
75%	4457.000000
max	35275.000000

Figure 8. Covid dataframe describe function

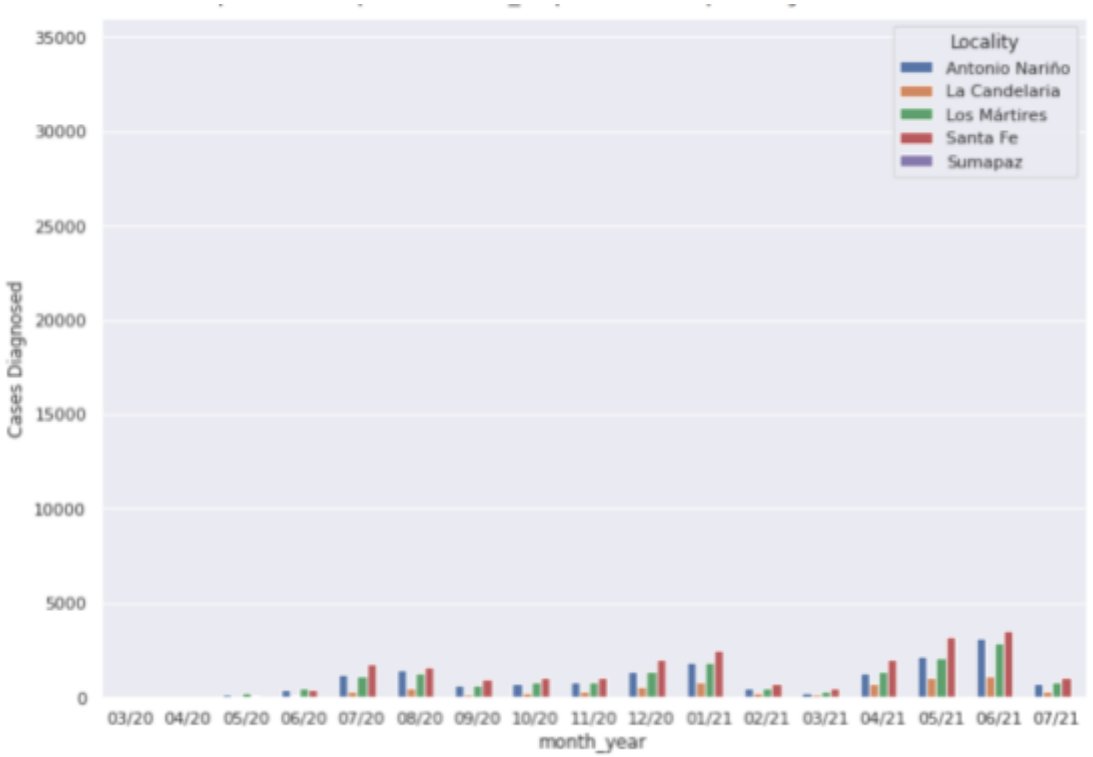
The average cases diagnosed per month is 3.627 and the maximum number of cases diagnosed in a month was 35.275. Since we are working with 20 different localities, it will be necessary to generate multiple graphs so as not to saturate on the same graph with multiple categories. We will divide the localities into groups of 5 and these groups will be selected by determining the order of average cases by localities.

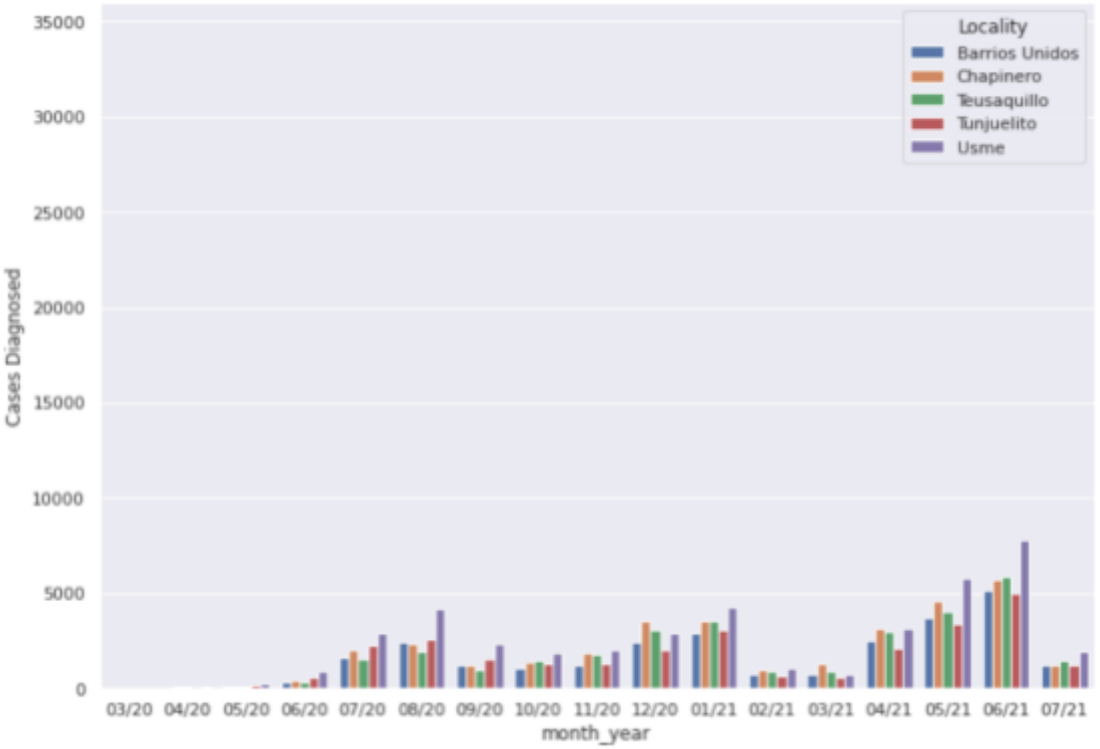
```
Locality
Sumapaz                      7.600000
La Candelaria              403.647059
Los Mártires               989.176471
Antonio Nariño             995.352941
Santa Fe                  1337.294118
Barrios Unidos            1611.000000
Tunjuelito                1636.294118
Teusaquillo               1828.000000
Chapinero                 1969.411765
Usme                      2490.235294
Puente Aranda             2883.058824
Rafael Uribe Uribe        3317.352941
San Cristóbal             3339.117647
Fontibón                  3786.647059
Ciudad Bolívar            4451.529412
Usaquén                   5442.058824
Bosa                      5542.764706
Engativá                  8388.529412
Kennedy                   9597.705882
Suba                     11025.117647
Name: Cases Diagnosed, dtype: float64
```

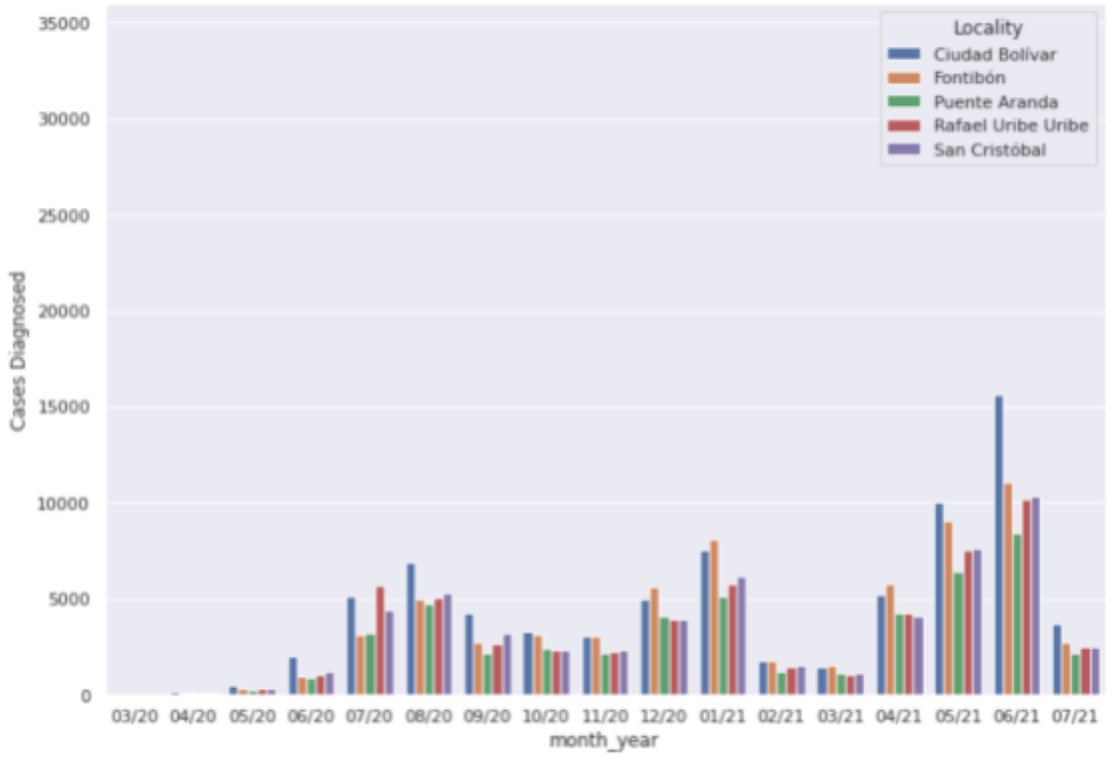
Figure 8. Average covid cases by locality (ordered)

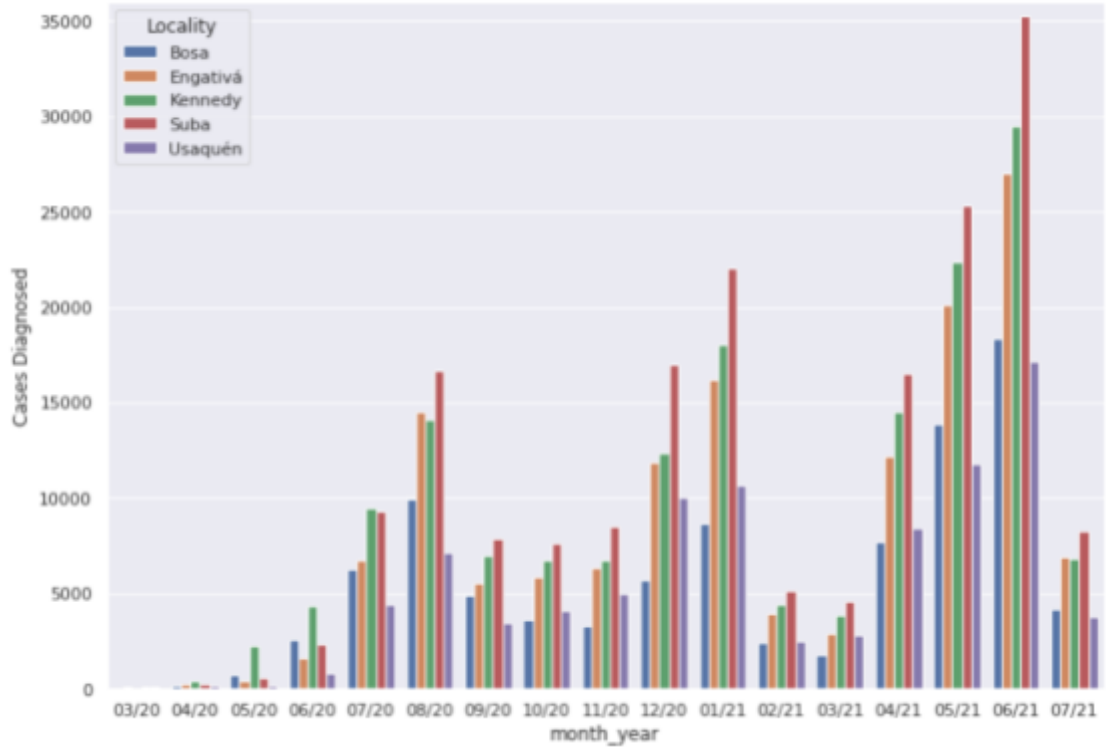
The graphs will be grouped as follows:

Group	Localities
1 (Lowest Average of Covid Cases)	Sumapaz, La Candelaria, Los Mártires, Antonio Nariño, Santa Fe
2	Barrios Unidos, Tunjuelito, Teusaquillo, Chapinero, Usme
3	Puente Aranda, Rafael Uribe Uribe, San Cristóbal, Fontibón. Ciudad Bolívar
4 (Highest Average of Covid Cases)	Usaquén, Bosa, Engativá, Kennedy, Suba


Locality
Antonio Nariño
La Candelaria
Los Mártires
Santa Fe
Sumapaz
Cases Diagnosed
35000
30000
25000
20000
15000
10000
5000
0
03/20
04/20
05/20
06/20
07/20
08/20
09/20
10/20
11/20
12/20
01/21
02/21
03/21
04/21
05/21
06/21
07/21
month_year


Locality
Barrios Unidos
Chapinero
Teusaquillo
Tunjuelito
Usme
Cases Diagnosed
35000
30000
25000
20000
15000
10000
5000
0
03/20
04/20
05/20
06/20
07/20
08/20
09/20
10/20
11/20
12/20
01/21
02/21
03/21
04/21
05/21
06/21
07/21
month_year





As it can be observed in the graphs, the localities with the most cases can even reach a difference of four or even five times the amount of cases in the localities which have the least number of cases. This might be an important observation to take into account later on when the best neighborhoods to open the new business are selected.

4. Analysis and Results

Since one of the project's objectives is to determine the similarity between neighborhoods so that they can be grouped into clusters with similar characteristics, one of the best techniques to apply is data clustering. The neighborhood and venues data has no predefined categories or groups, which is why we'll be using the K-Means Clustering algorithm.

This algorithm will divide the neighborhood data into non-overlapping subsets, this way we can make sure a neighborhood belong only to one category. In order to train this algorithm we will need two inputs: the k amount of clusters in which the neighborhoods will be categorized and the neighborhood data with its features.

As a first step, the features data that we use as an input for the K-Means algorithm need to be numerical variables. Our current venues data is stored as categorical variables, which is why we perform one-hot encoding to our venues data set (Specifically the Venue Category feature). We then proceed to group the data by neighborhood and calculate the mean of the frequency of each venue category.

In the second step step we will need to determine which value might be the best approach to the amount of k clusters we want the algorithm to return. This can be done by applying the 'elbow method'. This consists in calculating and plotting the variance between the amount of clusters and the mean distance of the data points in the cluster to its centroid.

We applied this method and the resulting graph was:

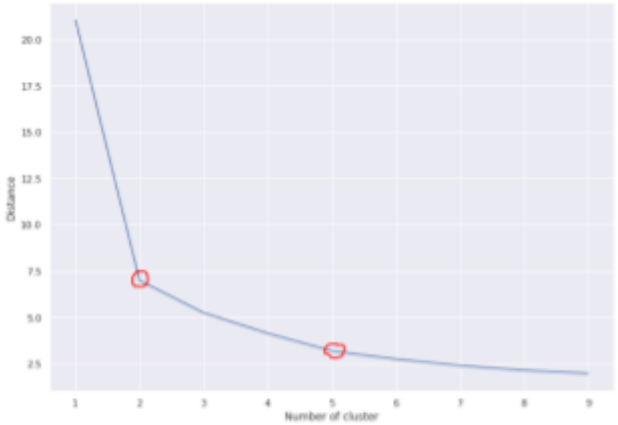


Figure 10. Elbow method for determining number of clusters

We can see that the first 'elbow' in the graph is when the number of clusters is set in 2. However, having only two clusters into which the neighborhoods were categorized doesn't give us an accurate view of which are the neighborhoods that might be candidates to set up the business. We go with the next value which is 5 clusters.

Once we have the amount of clusters and the features data, we run the algorithm and store the labels it returns for each neighborhood.

	Number	Neighborhood	Locality	Latitude	Longitude	Cluster Labels
1	9	Verbenal	usaquén	4.765150	-74.038394	2.0
2	10	La Uribe	usaquén	4.752400	-74.045013	2.0
3	11	San Cristóbal Norte	usaquén	4.734501	-74.017543	2.0
4	12	Toberín	usaquén	4.747274	-74.043719	3.0
5	13	Los Cedros	usaquén	4.720422	-74.116078	0.0

Figure 11. Preview of the dataframe with the neighborhoods and its cluster label

We then proceed to once again map the neighborhoods in Bogota, but this time each neighborhood will have a color representing the cluster it belongs to.

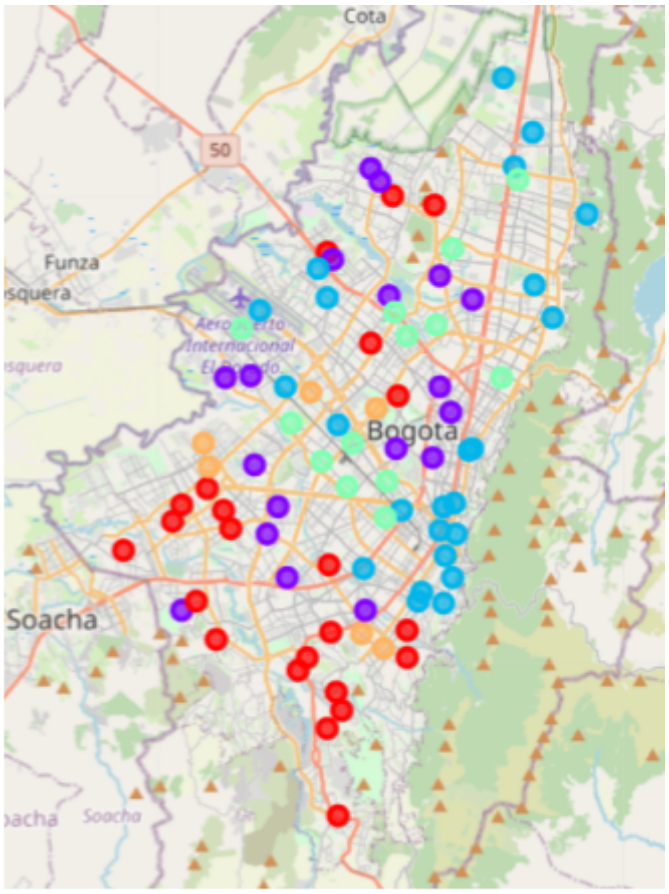


Figure 12. Neighborhoods Clusters

According to our last map and the clusterization of the neighborhoods, those neighborhoods that belong to cluster #1 (Cyan Cluster) are the ones who are most likely to have the most businesses related to books and cafes and that have parks in the vicinity.

With this information we will now proceed to check the amount of competition in case the user decides to open their business in one of these neighborhoods and check the amount of covid cases in the area in order to also take into account this information for the business location decision.

	Neighborhood	Coffee Shops	Book Shops	Parks	Locality	Cluster Labels
2	San Cristóbal Norte	1.0	NaN	NaN	usaquén	1.0
19	La Academia	1.0	NaN	NaN	suba	1.0
11	Las Cruces	2.0	NaN	1.0	santa fe	1.0
1	La Uribe	3.0	1.0	NaN	usaquén	1.0
15	Capellanía	3.0	1.0	1.0	fontibón	1.0
21	Santa Isabel	3.0	NaN	1.0	los mártires	1.0
13	Bavaria	6.0	NaN	NaN	kennedy	1.0
17	Garcés Navas	6.0	NaN	2.0	engativá	1.0
0	Verbenal	7.0	1.0	1.0	usaquén	1.0
4	Country Club	7.0	1.0	2.0	usaquén	1.0
14	Ciudad Salitre Occidente	7.0	1.0	2.0	fontibón	1.0
18	Álamos	7.0	NaN	1.0	engativá	1.0
12	Lourdes	11.0	NaN	NaN	santa fe	1.0

Figure 14. Preview of the table with the count information of business and parks

According to the information resumed in the table, in a top 5 of the best possible neighborhoods to start the new book café business are:

Neighborhood	Locality
Santa Isabel	Los Mártires
Garcés Navas	Engativá
Las Cruces	Santa fe
Capellanía	Fontibón
Verbenal	Usaquén

This is because these neighborhoods are part of the cluster of those neighborhoods that tend to have similar business such as coffee shops and bookstores. However, these locations do not have a large amount of these businesses so as to have a large amount of competition in the vicinity. Another of the factors taken into account was that the neighborhood had at least one park. If this factor is not of importance to our client these other neighborhoods can be considered:

Neighborhood	Locality
San Cristóbal Norte	Usaquén
La Academia	Suba

Having this information, we should now consider the behaviour in covid cases for the respective locality in each neighborhood. Similarly as with the data exploration we did at the beginning, we will once again graph the amount of covid cases per month, but this time we'll only graph the localities of our target neighborhoods.

Figure 15. Covid data graph for target localities

As we can see in the graph, the localities of Suba, Engativá and Usaquén seem to have had the most amount of covid cases since the beginning of the pandemic and it's behaviour has been similar in the past months of 2021. Taking this into account we can establish the next order of recommendation:

Pos	Neighborhood	Locality
#1	Santa Isabel	Los Mártires
#2	Las Cruces	Santafé
#3	Capellania	Fontibón
#4	Verbenal	Usaquén
#5	San Cristobal Norte	Usaquén
#6	Garcés Navas	Engativá
#7	La Academia	Suba

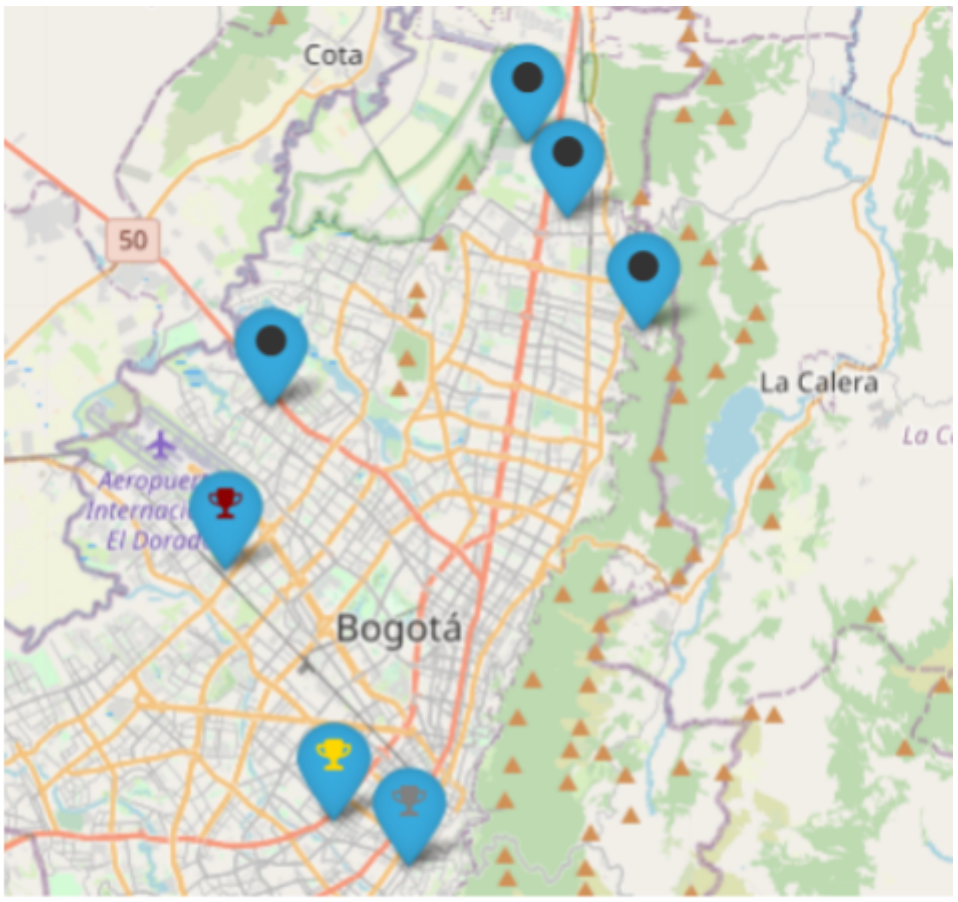


Figure 16. Final results

5. Conclusions and Observations

Taking into account the results, we can establish that the usage of the K-Means Clustering Algorithm is a good algorithm option to implement on these sorts of projects since we have information that can be considered as “characteristics”(venues) of our subjects (neighborhoods). By performing one-hot encoding into these characteristics, we have the necessary data to use k-means and determine groups of similar neighborhoods in Bogotá. By establishing which is the type of business our client would like to open (in this case a book café), the Foursquare API allows us to get filtered information on these categories in order to get a better result regarding which neighborhoods tend to have more of this kind of businesses.

One of the things we could observe from the information captured with the Foursquare API is that it does not return the full list of available venues. For example, in the Folium map we can observe there are some parks that are not listed by Foursquare. This also happens with some business venues. This could be due to the API version used or the category Ids that were sent into the API call. This is something that should be taken into account for future versions of the project or similar projects.

6. Future Steps

Later on, using this project as a reference, we could develop a new version in which the user is given the chance to select the sort of business they are interested in. Since we have the Foursquare API documentation, a list of available categories can be shown to the user and the id of each category can be used as an input for various of the functions that were created in this project. This would allow us a higher level of automation in the project and to make it available for even more customers.